音声分析方法および音声合成システム

発明の背景

この発明は、音声を分析してパラメータで表現し、それを圧縮／蓄積し、再び音声に合成する、いわゆる音声分析合成系に属する。

音声の分析合成系(speech analysis-synthesis system)はボコーダ(vocoder)と呼ばれ、音声信号をモデル化することによって少ないパラメータで効率良く音声信号を表現し、再び元の音声を合成するものである。音声を波形データのまま伝送するよりもはるかに少ないデータ量で伝送できるため、音声通信系で用いられている。その代表的なシステムはLPC(Linear prediction coding)分析合成系である。

しかし、LPC を始めとする多くのボコーダによって合成された音声では、肉声の持つ自然さが少なからず損ねられる。LPC ボコーダは有声音(voiced sound)の音源(sound source)および無声音(voiceless sound)の音源(sound source)としてそれぞれインパルス列(impulse series)および白色雑音(white noise)を仮定したモデルであるため、有声音部ではブザー的な(buzzy)音質になる。また、声帯(vocal tract)の振動波形がインパルス列とは異なるので、音源の持つスペクトル傾斜(spectral tilt)などの影響が正しく考慮されず、結果として声道伝達特性(vocal tract transfer characteristics)の推定誤差(estimation error)が大きくなる。

そこで、音源に声帯波形(glottal waveform)のモデルを用い、声道パラメータ(vocal tract parameter)と音源パラメータ(voice source parameter)とを同時に推定する方法が考案された。Dingらは、ARX(autoregressive-exogenous)音声生成モデルに基づくピッチ同期な音声分析合成方法(Ding, W., Kasuya, H., and Adachi, S., "Simultaneous Estimation of Vocal Tract and Voice Source Parameters Based on an ARX Model", IEICE Trans. Inf. & Syst., Vol. E78-D, No.6 June 1995)を開発した。しかしながら、この方法はピッチ周期が短い音声の分析や子音から母音への渡りの部分の分析に問題がある。

発明の概要

この発明の1つの局面に従うと、音声合成システムは、音声生成モデルに基づき推定されたフォルマントパラメータ(フォルマント周波数とフォルマントバンド幅とを含む)の時系列データを利用して音声を合成するものであり、隣り合ったフレーム間でのフォルマントパラメータの対応関係を動的計画法(dynamic programming)を用いて決定する。

好ましくは、上記音声合成システムは、フォルマントパラメータの対応関係の決定において、αとβを所定の重み係数、$F_f(n)$を第 n フレームのフォルマント周波数、$F_i(n)$を第 n フレームのフォルマント強度、ε を所定の値とするとき、

$$d_c(F(n),F(n+1)) = \alpha\left|F_f(n) - F_f(n+1)\right| + \beta\left|F_i(n) - F_i(n+1)\right|$$

$$d_d(F(k)) = \alpha\left|F_f(k) - F_f(k)\right| + \beta\left|F_i(k) - \varepsilon\right|$$

$$= \beta\left|F_i(k) - \varepsilon\right|$$

によって接続コスト$d_c(F(n), F(n+1))$および非接続コスト$d_d(F(k))$を求め、動的計画法における格子点移動のコストに用いる。

好ましくは、上記音声合成システムは、互いに接続されないフォルマントが含まれる隣り合ったフレームにおいて、接続される相手が存在しないフォルマントと同じ周波数で強度が 0 のフォルマントをもう一方のフレームに配置し、二つのフレームの間をフォルマント周波数と強度を滑らかな関数に従って補間して接続する。

好ましくは、上記音声合成システムは、$F_b(n)$を第nフレームのフォルマントバンド幅、F_sをサンプリング周波数とするとき、

$$F_i(n)=\begin{cases}20\log_{10}\left(\dfrac{1+e^{-\pi F_b(n)/F_s}}{1-e^{-\pi F_b(n)/F_s}}\right) & ,\text{if formant}\\ 20\log_{10}\left(\dfrac{1-e^{-\pi F_b(n)/F_s}}{1+e^{-\pi F_b(n)/F_s}}\right) & ,\text{if anti-formant}\end{cases}$$

によってフォルマント強度$F_i(n)$を計算する。

好ましくは、上記音声合成システムは、複数のフォルマントを含む声道伝達関数を複数のフィルタの縦続接続によって実現し、隣り合うフレームの間で接続が行われないフォルマントがあることによりフィルタの接続を変更する必要がある場合に、フィルタの係数および内部記憶値を別のフィルタにコピーし、自身のフィルタの係数および内部記憶値は別のフィルタからコピーするか所定の値に初期化する。

この発明のもう1つの局面に従うと、音声分析方法は、RK 音源モデルなどの声帯音源モデルを利用して音声信号波形の音源パラメータと声道パラメータを推定するものであり、推定された声道伝達関数の逆特性にて構成されたフィルタを用いて推定音源波形を抽出し、前記推定音源波形の声門閉鎖タイミング(GCI: glottal closure instance)に対応するピーク位置を2次関数などのあてはめによってサンプリング周期よりも高い時間精度で推定し、前記推定されたピーク位置の近傍のサンプル位置に GCI を同期させて音源モデル波形を生成し、前記生成された音源モデル波形をオールパスフィルタでサンプル周期よりも高い時間精度で時間的に移動することによりGCIを前記推定されたピーク位置に一致させる。

この発明のさらにもう1つの局面に従うと、音声分析方法は、RK 音源モデルまたはその拡張として定義される声帯音源モデルを利用して音声信号波形の音源パラメータと声道パラメータを推定するものであり、推定された声道伝達関数の逆特性にて構成されたフィルタを用いて推定音源波形を抽出し、前記推定音源波形の DFT(discrete Fourier transformation)における基本波レベルをH1、第2高調波レベルを H2 として、HD=H2-H1 で定義される HD の値から声門開放時間率(open quotient) OQ を推定する。

好ましくは、上記音声分析方法では、OQ の推定に、

$OQ = 3.65HD - 0.273HD^2 + 0.0224HD^3 + 50.7$ の関係を用いる。

図面の簡単な説明

図1は、ARX 音声生成モデルを示すブロック図である。

図2は、RKモデルのOQパラメータ値と第1高調波、第2高調波レベル差の関係を表した図である。

図3は、オールパスフィルタを用いた場合の音源パルス波形の例を示す図であり、(a)は原波形、(b)はT_d=50μsでシフトされた波形、(c)はd_g=3msのランダマイズとシフトが施された波形である。

図4(a)は不連続なフォルマント、図4(b)はそのスペクトル変化の様子を示す図である。

図5は、聴取実験による評価結果を示す図である。

図6は、この発明の第1の実施形態による音声分析システムの構成を示すブロック図である。

図7は、音声分析処理の流れを説明するための図である。

図8は、AVパラメータの求め方を説明するための図である。

図9は、複素数の極座標表示の概念を表した図である。

図10は、GCIの高精度推定の説明図である。

図11(a),(b)は、RKモデル音源波形をオールパスフィルタでサンプル周期よりも細かい精度でシフトする方法を説明するための図である。

図12は、この発明の第3の実施形態による音声合成システムの構成を示すブロック図である。

図13は、この発明の第4の実施形態による音声合成システムにおけるRKモデル音源生成部の構成を示すブロック図である。

図14は、この発明の第5の実施形態による音声合成システムの構成を示すブロック図である。

図15は、隣り合った2つのフォルマントのフォルマント周波数とバンド幅の関係を示す図である。

図16は、横軸にFrame A、縦軸にFrame Bのフォルマントをとった格子の概念を説明するための図である。

図17は、すべてのフォルマントが同じ番号同士接続された場合の格子の様子を説明するための図である。

図18は、接続されないフォルマントがある場合の格子の様子を説明するための図である。

図19は、移動の制約を説明するための図である。

図20は、図19の制約のもとに通り得る格子点を示す図である。

図21は、パス探索の流れを説明するための図である。

図22は、パス探索によって計算されたコストの例を説明するための図である。

図23は、パスBが選択された様子を示す図である。

図24は、求められた最適パスを示す図である。

図25は、最適パスに従って接続されたフォルマントの様子を示す図である。

図26は、Frame AとFrame Bの周辺を拡大した図である。

図27は、接続先のないフォルマントのために相手側のフレームに強度0のフォルマントを配置した様子を示す図である。

図28(a),(b)は、フォルマントフィルタの構成を示す図である。

図29は、フォルマントフィルタの縦続接続構成の変更方法を説明するための図である。

図30は、フォルマントフィルタの縦続接続構成を変更する処理の流れを示す図である。

好ましい実施形態の説明

まず、ARX(autoregressive-exogenous)音声生成モデルに基づく音声分析合成方法の概要について説明する。

[ARX 音声生成モデル]

ARX 音声生成モデルは図1に示すとおりであり、また、以下の式(1)に示す線形差分方程式によって表される。

$$y(n)+\sum_{k=1}^{p}a_k y(n-k)=\sum_{k=0}^{q}b_k u(n-k)+e(n) \qquad (1)$$

ここで入力$u(n)$は周期的な音源波形を表し、出力$y(n)$は音声信号を表す。声帯のノイズ成分は白色雑音$e(n)$によってシミュレートされている。式(1)において、a_iおよびb_iは声道フィルタ係数(vocal tract filter coefficients)であり、pおよびqはARXモデルの次数である。

ここで$A(z)$および$B(z)$を以下のように定義する。

$$A(z)=1+a_1z^{-1}+\cdots+a_pz^{-p}$$

$$B(z)=b_0+b_1z^{-1}+\cdots b_qz^{-q}$$

すると式(1)のz変換は次のように表される。

$$Y(z)=\frac{B(z)}{A(z)}U(z)+\frac{1}{A(z)}E(z) \qquad (2)$$

ここで$Y(z),U(z)$および$E(z)$はそれぞれ$y(n),u(n)$および$e(n)$のz変換である。声道の伝達関数は$B(z)/A(z)$で与えられる。

放射特性(radiation characteristics)を含んだ微分声門体積流(differentiated glottal flow waveform)を表すためにRK(Rosenberg-Klatt)モデル(Klatt, D. and Klatt, L., "Analysis synthesis and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Amer. Vol. 87, 820-857,1990)を使用する。RK波形は次のように表される。

$$rk(n)=rk_c(nT_s) \qquad (3)$$

$$rk_c(t)=\begin{cases}2at-3bt^2, & 0\le t<OQT0\\ 0, & elsewhere\end{cases} \qquad (4)$$

$$a=\frac{27AV}{4OQ^2T0},\ b=\frac{27AV}{4OQ^3T0^2}$$

ここでT_sはサンプリング周期、AVは振幅パラメータ、$T0$はピッチ周期、OQは声門開放率(an open quotient of the glottal open phase of the pitch period)である。微分声門体積流$u(n)$は$rk(n)$をローパスフィルタによりスムージングして得られる。スペクトル包絡の傾斜はスペクトル傾斜パラメータTLによって調整される。ローパスフィルタは次のように定義される

$$TL(z)=(1-cz^{-1})^{-2} \qquad (5)$$

ローパスフィルタ係数cは以下の式(6)によって傾斜パラメータTLに関連づけられる。

$$TL = 20\log_{10}\left|TL(e^{j0})\right| - 20\log_{10}\left|TL(e^{j\omega_0})\right|,$$
$$c = \frac{B - \cos\omega_0 - \sqrt{(B-\cos\omega_0)^2 - (B-1)^2}}{B-1} \tag{6}$$

ここで、$B = 10^{TL/20}, \omega_0 = 2\pi 3000/F_s$ である。

[分析アルゴリズム]

[フィルタ係数の推定]

Ding らはカルマンフィルタアルゴリズムを使用し、調音移動(articulatory movement)を考慮に入れてARXモデルの時変係数を逐一推定するが、ほとんどの場合1ピッチ周期あたり1セットの係数だけを蓄える形式をとっていた。2,000Hz 以下のバンド幅を有するすべてのフォルマント値を平均化することによって係数セットが獲得される。しかし、バンド幅の広いフォルマントが計算時に除かれるときには平均係数は適切なものにはならない。ここでは、分析フレームにわたって平均化された係数を推定する代わりに簡単な最小二乗法を使用する。

φ および θ を次のように定義する。

$$\varphi(n) = \left[-y(n-1)\cdots - y(n-p)u(n)\cdots u(n-q)\right]^T,$$
$$\theta = \left[a_1 \cdots a_p b_0 \cdots b_q\right]^T$$

すると式(1)は次のように表すことができる。

$$y(n) = \varphi^T(n)\theta + e(n), \quad n = 1,\cdots,N \tag{7}$$

予測誤差は次のようになる。

$$\varepsilon(n,\theta) = y(n) - \varphi^T(n)\theta \tag{8}$$

最小二乗基準関数は次のようになる。

$$V(\theta) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\varepsilon^2(n,\theta) \tag{9}$$

最小二乗推定値は以下のように与えられる。

$$\hat{\theta} = \arg\min_{\theta} V(\theta)$$
$$= \left[\frac{1}{N}\sum_{n=1}^{N}\varphi(n)\varphi^T(n)\right]^{-1}\frac{1}{N}\sum_{n=1}^{N}\varphi(n)y(n) \tag{10}$$

[スペクトル傾斜の補償]

$A(z)$ および $B(z)$ の実軸上にある根およびバンド幅がとても広い根は声道の共振に関連しないため除外する必要がある。しかしそれらの根を単に除外することにより声道伝達関数のスペクトル傾斜が変わる。以下に示す $C(z)$ のスペクトル傾斜を補償するためにシステム伝達関数 $D(z)$ を導入する。

$$C(z) = \frac{B(z)}{A(z)}\frac{A'(z)}{B'(z)} \tag{11}$$

ここで $B'(z)/A'(z)$ は除外されないフォルマントからなる。$C(z)$ のスペクトル傾斜を近似するため、実軸上の2重極または零点の $D(z)$ を次のように定義する。

$$D(z) = (1 - dz^{-1})^{\mathrm{sgn}(TL)2} \tag{12}$$

ここでsgn(·)は()内の値の符号を表す。スペクトル傾斜パラメータTIは次のように与えられる。

$$TI = 20\log_{10}\left|C(e^{j0})\right| - 20\log_{10}\left|C(e^{j\omega_0})\right|$$

ここで$\omega_0 = 2\pi 3000/F_s$である。式(12)中の係数dは式(6)と同様にしてTLから導出される。

[音源の生成]

フォルマントの推定をより安定して行うため任意の長さのマルチパルス音源を生成する。マルチパルス音源信号$v(n)$は次のように与えられる。

$$v(n) = \sum_{i=1}^{M} rk(n - OQ\,T0\,F_s + GCI(i), AV(i), T0, OQ) \tag{13}$$

$T0$は分析フレームにおけるピッチ周期の平均値である。OQの初期値は適当な値に設定される。音声振幅パラメータ$AV(i)$および声門閉鎖時刻$GCI(i)$は、逆フィルタリングされた音声$v'(n)$の駆動ピーク(excitation peaks)より得られる。逆フィルタ波形$v'(n)$のz変換は次のように与えられる。

$$V'(z) = \left(\frac{A'(1)}{B'(1)D(1)TL(1)}\right)^{-1} \frac{A'(z)}{B'(z)D(z)TL(z)} Y(z) \tag{14}$$

$v'(n)$の駆動振幅(excitation amplitude)AEはAVパラメータに変換される。

$$AV = \frac{4}{27} OQ \cdot AE \tag{15}$$

[適応プリフィルタ]

パーシバルの定理を用いると次のように式(9)を周波数領域で表すことができる(Ljung, L., "System identification theory for the user." PRENTICE HALL PTR, 201-202, 1995)。

$$V(\theta) \cong \frac{1}{2N} \sum_{k=0}^{N-1} \left\{ \left| G(e^{j2\pi k/N}) - \frac{B(e^{j\frac{2\pi}{N}k}, \theta)}{A(e^{j\frac{2\pi}{N}k}, \theta)} \right|^2 \left| W(\frac{2\pi}{N}k, \theta) \right|^2 \right\} \tag{16}$$

ここで、

$$G(e^{j2\pi k/N}) = \frac{Y(\frac{2\pi}{N}k)}{U(\frac{2\pi}{N}k)},$$

$$Y(\frac{2\pi}{N}k) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} y(n) e^{-j2\pi k/N}, \tag{17}$$

$$U(\frac{2\pi}{N}k) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} u(n) e^{-j2\pi k/N},$$

$$W(\omega, \theta) = U(\omega) A(e^{j\omega}, \theta)$$

である。

式(16)から、予測誤差法は経験的伝達関数推定値(ETFE)$G(e^{j2\pi k/N})$に重みづけ関数$W(\omega, \theta)$を用いてモデル声道伝達関数を適合させる方法であると説明することができる。

システムの入力信号および出力信号が以下に示す$L(z)$でプリフィルタリングされているとする。

$$L(z) = 1 + l_1 z^{-1} + l_2 z^{-2} \cdots + l_r z^{-r} \tag{18}$$

すると重みづけ関数は次のように表すことができる。

$$W(\omega,\theta) = U(\omega)A(e^{j\omega},\theta)L(e^{j\omega}) \tag{19}$$

このことは、プリフィルタ$L(z)$によって$W(\omega,\theta)$を制御できることを意味する。ARX 音声生成モデルでは、音源$U(\omega)$のスペクトル傾斜はTLによって決定され、$A(e^{j\omega})$は局所的な周波数レンジにおいて反共振を持つけれども$A(e^{j\omega})$のスペクトル傾斜は広い周波数レンジにおいてフラットであると仮定される。Ding らはスペクトル傾斜パラメータTLの影響を無視して$L(z) = 1 - z^{-1}$で表される不変フィルタを使用した。

ここでは、重みづけ関数$W(\omega)$における$U(\omega)$をキャンセルするため、TLの影響を考慮した適応プリフィルタ$L(z)$を使用する。プリフィルタ$L(z)$の係数は、LS 法を使用した次の AR モデルより得られる。

$$u(n) = \sum_{k=1}^{r} l_k u(n-k) + \xi(n) \tag{20}$$

ここでモデルの次数はrであり、典型的には6または 8 である。$\xi(n)$は白色雑音である。

[声門開放率の推定]

図2に示すようにRKモデルの声門開放率OQはマルチパルス音源の第1高調波レベル($H1$)と第2の高調波レベル($H2$)とに強く関連している。OQ [%]は次の式(21)によって与えられる。

$$\begin{aligned} OQ &= 3.65HD - 0.273HD^2 + 0.0224HD^3 + 50.7, \\ &-4.03 \le HD \le 9.83 \end{aligned} \tag{21}$$

ここで、$HD = H2 - H1$[dB]である。$H2$および$H1$は、式(14)によって逆フィルタされた音声のDFTから得られる。

[合成アルゴリズム]

有声音および無声音の双方を合成するために、縦続接続されたフォルマント合成器を使用する。RK モデルを使用して有声音を合成し、M 系列による擬似白色2値信号を使用して無声音を合成する。

[音源の制御]

以下の2つの問題を解決するために、RK 音源に対し2つのオールパスフィルタ(APF)(Kawahara, H., "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited', ICASSP 97, 1303-1306, 1997)を適用する。

- 連続する2つの声門閉鎖(GCI)の間隔が$F0$の認知のcueになっていると考えられるので,RKモデルの声門閉鎖時刻を適切に制御しなければならない。
- コンスタントな音源信号の繰り返しがバジーな音質を引き起こしていると考えられるので、何らかの適切な揺らぎを音源信号に加える必要がある。

改良された音源$rk'(n)$は次の式に従う。

$$\begin{aligned} rk'(n) &= \frac{1}{\sqrt{N}} \sum_{k=-N/2+1}^{N/2} R'(\frac{2\pi}{N}k) e^{j2\pi kn/N} \\ R'(\frac{2\pi}{N}k) &= R(\frac{2\pi}{N}k) e^{j(\Theta_r(k) - \Theta_r(k))} \end{aligned} \tag{22}$$

ここで$R(2\pi k/N)$は式(3)のDFTである。

$$R(\frac{2\pi}{N}k)=\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}rk(n)e^{-j2\pi k/N} \tag{23}$$

位相$\Theta_s(k)$は音源波形をT_d [sec]だけシフトさせる。

$$\Theta_s(k)=\frac{2\pi}{N}\frac{T_d}{F_s}k \tag{24}$$

一方、$\Theta_r(k)$は高周波数領域の群遅延をランダム化する。

$$\Theta_r(k)=\begin{cases}\eta'(k), & k=0,\cdots,\dfrac{N}{2}\\ -\eta'(-k), & k=-\dfrac{N}{2}+1,\cdots,-1\end{cases}$$

$$\eta'(k)=\frac{2\pi}{N}\sum_{l=0}^{k}w_\eta(l)\eta(l) \tag{25}$$

$$w_\eta(l)=\frac{1}{1+e^{(w_c-2\pi l/N)/w_t}}$$

$$\eta(l)\sim N(0,d_gF_s),\quad l=0,\cdots,\frac{N}{2}$$

群遅延$\eta(l)$は平均ゼロおよび分散d_gF_s [point]の白色雑音である。カットオフ周波数ω_c [rad](典型的には$2\pi 100/F_s$)によって定義される高周波数における位相を重みづけ窓$w_\eta(l)$を使用して操作する。一例を図3に示す。

[最適なフォルマント接続]

上述の自動推定は、声道伝達関数の係数が連続的に変化するということを必ずしも保証しない。時変システムであるフォルマント合成器では、デジタルフィルタ係数の不連続点はクリック音を生じさせる。不連続点は次の2つのケースで生じる。1)2つの連続するフレーム間でのフォルマントの数が同じではないとき、2)フォルマント周波数が急激に変化するときである。

動的計画法を適用して接続コスト$d_c(F(n),F(n+1))$および非接続コスト$d_d(F(k))$からなる距離尺度を用いたフォルマント$F(n)$とフォルマント$F(n+1)$との間の最適なマッチングを得る。

$$d_c(F(n),F(n+1))=\alpha\left|F_f(n)-F_f(n+1)\right|+\beta\left|F_i(n)-F_i(n+1)\right| \tag{26}$$

$$\begin{aligned}d_d(F(k))&=\alpha\left|F_f(k)-F_f(k)\right|+\beta\left|F_i(k)-\varepsilon\right|\\&=\beta\left|F_i(k)-\varepsilon\right|\end{aligned} \tag{27}$$

ここでF_fはフォルマント周波数であり、F_iはフォルマント強度である。フォルマント強度F_iは、フォルマントスペクトルの最大レベルと最小レベルとの差として定義される。

$$F_i(n)=\begin{cases}20\log_{10}\left(\dfrac{1+e^{-\pi F_b(n)/F_s}}{1-e^{-\pi F_b(n)/F_s}}\right) & \text{, if formant}\\ 20\log_{10}\left(\dfrac{1-e^{-\pi F_b(n)/F_s}}{1+e^{-\pi F_b(n)/F_s}}\right) & \text{, if anti}-\text{formant}\end{cases} \tag{28}$$

フォルマントが対応する相手を持たないとき、当該フォルマントと同じ周波数および非常に小さい

強度εを持つフォルマントが接続されるべきフォルマントであるとみなされる。フォルマント最適接続のシミュレーション結果を図4に示す。図4に示すように、フォルマント周波数が急に変化してもスペクトル包絡は非常にスムーズに変化する。

[実験]

18人の男性および5人の女性によって読まれた日本語のある長文で分析合成システムの実験を行った。18人の男性話者は、話者の個人性に関連する音質の変化を研究するために準備された108人の男性の音声データ集合から選択され、音質変化に関してオリジナルの 108 人の男性を十分に表しているとみなされている(Ljung, L, "System Identification theory for the user." PRENTICE HALL PTR, 201-202, 1995)。

提案した方法のある方法に対する合成音声の質の点での優越性が Ding らによって確認された後、周知のメルケプストラム(MCEP)法(Tokuda, K., Matsumura, H., and Kobayashi, T. "Speech coding based on adaptive mel-cepstral analysis." ICASSP 94, 197-200, 1994)と ARX 法との間でさらに比較を行った。以前の実験で使用されたのと同じ音声サンプルを用いた。デジタル化のためのサンプリング周波数を 11.025kHz とした。ピッチ変換に対する頑健さをテストするため、再合成によってオリジナルよりも1．5倍高い基本周波数を持つものも音声サンプルとした。より自然に聞こえるほうの合成音声を選択するように5人の被験者に頼むことによって比較テストを行った。結果を図5に示す。図5では、低いピッチおよび高いピッチの2つのピッチグループとピッチ変換とについての統計を示している。低いピッチの音声データについては ARX 法と MCEP 法との差は小さいけれども、高いピッチの音声については ARX 法のほうがよりよく機能することは明らかである。

(第1の実施形態)

図6はこの発明の第1の実施形態による音声分析システムの構成を示す図である。このシステムの動作は図7に示す流れに沿って行われる。以下、図6および図7を参照しつつ説明する。

まず、分析対象の音声波形から、25msec～35msec 程度の長さの窓関数で音声波形(speech segment)601を切り出す。窓関数としてはよく知られたHanning 窓などを用いる。25msec～35msec の長さは従来の分析方法と比較するとかなり長いものであり、男女の通常のピッチ範囲の音声波形から数ピッチ周期をまとめて切り出す程度の長さである。

<ステップ S7001>

次に GCI 検出部602および AV 推定部603において、音声波形601から負方向のピークピッキングを行い GCI(glottal closure instance)およびAV の初期値を求める。GCI には音声波形601の負方向のピーク位置を用いる。AV は、ピークの値が RK 音源波形の負のピークと一致するような値として、式(15)および図8に示すように求める。

<ステップ S7002>

次に音源波形生成部604において、図1および式(3)に示した RK モデル波形を、その負のピーク位置が GCI に同期するように発生させ、音源波形605とする。このときのパラメータは、AV にはステップ S7001 で求めた値、OQ の初期値であるOQ^0 には 0.6、TL^0 には 5～15 の値を適当に定めて用いる。$T0$ は分析対象フレームにおける平均ピッチ周期である。音源波形生成部604は式(13)に

従って音源波形605を生成する。

＜ステップ S7003＞

次にAR分析部606において、音源波形生成部604によって生成された音源波形605をAR分析する。AR分析モデルの次数は6または8を用いる。そして適応プリエンファシスフィルタ607および608は、AR分析の結果得られたフィルタ係数で音源波形605および音声波形601に対して適応プリエンファシス(adaptive pre-emphasis)を行う。適応プリエンファシスフィルタ607および608は式(18)で表される。

＜ステップ S7004＞

次にARX分析部609において、適応プリエンファシスフィルタ607および608によって適応プリエンファシスされた音源波形605および音声波形601を用いて式(7)から式(10)に示した方法でARX分析を行う。その結果、式(10)からAR係数a_iおよびMA係数b_iが求められ、式(2)における A(z)および B(z)が決定する。この時、A(z)=0、B(z)=0 と置いた方程式を解くことによりフォルマント周波数(formant frequency)$F_f(n)$、バンド幅(band-width)$F_b(n)$、アンチフォルマント(anti-formant)周波数$AF_f(n)$、バンド幅$AF_b(n)$を求める。すなわち、A(z)=0 の複素数解を$r_1,\cdots r_p$、B(z)=0 の複素数解を$s_1,\cdots,s_q$とすると、

$$F(k)=\left|\frac{\arg r_k}{2\pi T_s}\right|,\quad B(k)=\left|\frac{\ln|r_k|}{\pi T_s}\right|$$

$$AF(l)=\left|\frac{\arg s_l}{2\pi T_s}\right|,\quad AB(l)=\left|\frac{\ln|s_l|}{\pi T_s}\right|$$

によって求めることができる。ここで、

$$\arg c=\arctan\frac{\mathrm{Im}(c)}{\mathrm{Re}(c)}$$

$$|c|=\sqrt{\mathrm{Re}^2(c)+\mathrm{Im}^2(c)}$$

である。すなわち図9に示すように複素数 c を極座標で表現したものである。

なお、この時、バンド幅が大きいものはフォルマントパラメータから除外する。その結果、推定されるスペクトルの傾斜(spectral tilt)が影響を受けるため、式(11)から式(12)に示した方法でTIを推定する。

＜ステップ S7005＞

次に、推定されたフォルマントパラメータ$F_f(n)$, $F_b(n)$、スペクトル傾斜TI、音源スペクトル傾斜TL^0を用いて式(14)に示す逆フィルタ610を構成し、音声波形601から音源波形611を推定する。

＜ステップ S7006＞

次にOQ推定部612においてOQが推定される。具体的には、逆フィルタ610によって推定された音源波形611の DFT(discrete Fourier transform)から第1高調波のレベル$H1$と第2高調波のレベ

ル $H2$ の差である $HD = H2 - H1$ を求め、式(21)を用いて推定する。

＜ステップ S7007＞

次に GCI 検出部602および AV 推定部603において、逆フィルタ610によって推定された音源波形611から負方向のピークピッキングを行い GCI および AV の値が求められる。GCI および AV の求め方はステップ S7001 と同様である。

＜ステップ S7008＞

次に判定部613において、GCI が一定値に収束しているかどうかが判定される。収束していなければステップ S7002 から再び推定を繰り返す。収束していれば分析は終了し、次のフレームの分析へと移行する。なおフレームの周期は 5ms～10ms 程度が望ましい。

以上のように第1の実施形態による音声分析システムでは、分析窓長を25msec～35msec程度と従来よりも長くとり、複数のピッチ周期に対する音源位置を一括して推定することなどにより、高いピッチ周波数(pitch frequency)の女性の音声などに対してきわめて精度よく音源パラメータと声道伝達関数を推定することができる。

(第2の実施形態)

第1の実施形態では、ステップ S7001 および S7007 におけるピークピッキングによる GCI 推定は1サンプル単位で行われる。第2の実施形態では、これを1サンプル以下の精細な時間精度で行い、かつステップ S7002 において GCI に高精度に同期した RK 音源波形を生成することにより、分析精度を向上させる。

GCI の高精度推定の方法を図10に示す。音声波形601または逆フィルタ610によって推定された音源波形611の負方向のピークの位置を 2 次補間により精密に求める。すなわち、サンプル単位でピーク8001を検出し、その前後のサンプル8002および8003をあわせた3点を通る二次関数8004を求め、二次関数 8004 のピーク位置 8005 とピーク値 8006 を求める。

このようにして求められたピーク位置 8005 が GCI となるが、その値は整数のサンプル位置ではなく、実数値となる。そこで、RK 音源モデルの負のピーク位置を実数の GCI 位置に合わせるためにオールパスフィルタによる時間移動を行う。すなわちひとつのピッチ周期に対応する RK 音源モデルを式(22)－式(24)に従って移動する。ただし、ここでは $\Theta_r(k) = 0$ である。図10に示した推定されたピーク位置 8005 とその直前のサンプルの位置 8002 との時間差を式(24)の T_d に代入すればよい。

図11(a)は RK モデル音源波形をオールパスフィルタによりサンプル周期よりも細かい精度でシフトする例を表したものである。図11(b)に示すグラフには、もとの RK 音源波形、0.5 ポイントシフトしたもの、0.9ポイントシフトしたものを重ねて描いてある。このように、サンプル周期よりも細かい精度で GCI を同期させることによって、分析精度を向上することができる。

以上のように第2の実施形態による音声分析システムでは、音源の位置の推定において、音声波形あるいは逆フィルタによって推定された音源波形の負方向のピーク位置を2次補間により精密に求め、その位置にRK音源モデルの負方向のピーク位置が合うようにオールパスフィルタによる時間

移動を行う。このことにより、高い精度でGCIの推定が可能となり、音源パラメータ、声道伝達関数の推定精度を高めることができる。

（第3の実施形態）

図12は、この発明の第3の実施形態による音声合成システムの構成を示すブロック図である。この音声合成システムは、式(2)に従って合成音声を生成するものであり、RKモデル音源生成部12001、音源スペクトル傾斜フィルタ(source spectral tilt filter)($TL(z)$)12002、声道スペクトル傾斜フィルタ(Vocal tract spectral tilt filter)($D(z)$)12003、声道フィルタ(Vocal tract filter)($B(z)/A(z)$)12004、白色雑音生成部12005、白色雑音フィルタ($1/A(z)$)12006、混合部12007を備える。

第1または第2の実施形態に示した音声分析システムによって分析された音声は分析フレームごとに以下のパラメータに表現され、この音声合成システムに伝達される。

パラメータの種類	パラメータ名	意味
音源パラメータ	AV	RKモデル音源の振幅
	OQ	RKモデル音源の声門開放率
	F0	RKモデル音源の基本周波数
	TL	スペクトル傾斜量
	NA	白色雑音振幅
スペクトル傾斜補償フィルタ	TI	スペクトル傾斜補償量
フォルマント	F1～F6	第1～第6フォルマント中心周波数
	B1～B6	第1～第6フォルマントバンド幅

ここで、AV,OQ,F0,TLは有声区間(voiced part)のみで値を持ち、無声区間(voiceless part)では0である。一方、NAは無声区間のみで値を持ち、有声区間では0である。

RKモデル音源生成部12001は、パラメータAV, OQ, F0 を用いて式(13)に従って音源波形を生成する。音源スペクトル傾斜フィルタ12002は、パラメータTLを用いて式(5)に従って音源生成部12001からの音源波形のスペクトル傾斜を変更する。声道スペクトル傾斜フィルタ12003は、パラメータTIを用いて式(12)に従ってスペクトル傾斜を補償する。声道スペクトル傾斜フィルタ12003によってスペクトル傾斜が補償された音源波形は声道フィルタ12004を通って混合部12007に与えられる。すなわち式(2)の右辺第1項に従った音源波形が混合部12007に与えられる。白色雑音生成部12005は、パラメータNAに応じたゲインのランダム雑音を生成する。白色雑音生成部12005からのランダム雑音は白色雑音フィルタ12006を通って混合部12007に与えられる。すなわち式(2)の右辺第2項に従った雑音波形が混合部12007に与えられる。混合部12007は、声道フィルタ12004からの音源波形と白色雑音フィルタ12006からの雑音波形とを合成して式(2)に従った合成音声を生成する。

以上のように第3の実施形態による音声合成システムによれば、第1または第2の実施形態による音声分析システムによって推定されたパラメータを分析フレームごとに合成することにより、きわめて原音に近い高品質な音声を合成することができる。

（第4の実施形態）

この発明の第4の実施形態による音声合成システムは、図12に示したRKモデル音源生成部12001に代えて図13に示すRKモデル音源生成部13001を備える。その他の構成は図12に示した音声合成システムと同様である。図13に示すRKモデル音源生成部13001は、RKモデル音源生成部12001と、DFT(discrete Fourier transformation)計算部13002と、DFT変形部13003と、IDFT(inverse discrete Fourier transformation)計算部13004と、定常遅延量計算部13005と、ランダム遅延量計算部13006と、合成部13007とを備える。

RKモデル音源生成部12001は図12に示したものと等価なものである。DFT計算部13002は、RKモデル音源生成部12001からの音源波形を式(23)に従ってDFTして周波数領域(frequency domain)に変換する。定常遅延量計算部13005は、パラメータF0を用いて式(24)に従って定常遅延量$\Theta_s(k)$を計算する。ランダム遅延量計算部13006は式(25)に従ってランダム遅延量$\Theta_r(k)$を計算する。合成部13007は、定常遅延量$\Theta_s(k)$とランダム遅延量$\Theta_r(k)$とを加算($\Theta_s(k)-\Theta_r(k)$)してDFT変形部13003に与える。DFT変形部13003は、DFT計算部13002からの周波数領域の音源波形を式(22)の第2式に従って変形する。IDFT計算部13004は、DFT変形部13003によって変形された周波数領域の音源波形を式(22)の第1式に従ってIDFTして時間領域に戻す。

以上のようにして音源波形に揺らぎを付加することにより、

1)精密な声門閉鎖(glottal closure)タイミングの制御

2)バジー(buzzy)な音質の抑制

が可能となる。

（第5の実施形態）

図14は、この発明の第5の実施形態による音声合成システムの構成を示すブロック図である。図14に示す音声合成システムは、図12に示した音声合成システムの構成に加えてさらにフォルマント接続部14001を備える。フォルマント接続部14001は、隣接フレーム間のフォルマントパラメータF1～F6, B1～B6の連続性を考慮し、最適に接続するものである。フォルマント接続部14001は、式(26)および(27)に示した接続コスト(connection cost)および非接続コスト(disconnection cost)を用いて動的計画法(dynamic programming)によってフレーム間のフォルマントの対応を求める。

動的計画法の動作について以下に詳しく述べる。

図15は、ある二つの隣り合ったフレームのフォルマント周波数(formant frequency)とバンド幅(band width)を表している。横軸はフレーム番号で縦軸は周波数である。また、各フォルマントは(周波数,バンド幅)のように値を表示している。二つのフレーム(Frame AとFrame B)にはともに6個のフォルマントがある。これらのフォルマントは周波数の低いものから F1、F2・・・というように呼ばれる。通常、これら6つはFrame AとFrame Bの間で同じ番号のもの同士が接続される。しかし、Frame BのF2とF3の周波数は接近しており、どちらもFrame AのF2の周波数に近い。また、Frame BのF2はバンド幅が非常に大きい値となっている。バンド幅が大きいフォルマントは強度が弱く、消えつつあるか

出現しつつあるものと考えられる。したがって、Frame B の F2 は出現しつつあるフォルマントとみなされ、Frame A の F2 とは接続しないことが望ましい。Frame A の F2 は Frame B の F3 と接続されるべきである。このようなことを自動的に判断するために動的計画法を用いる。

図16は、横軸に Frame A、縦軸に Frame B のフォルマントをとり、格子状の点に(1, 1), (1,2)というように番号を振ったものである。この図では、フォルマントのそれぞれについて(周波数, 強度)というように値を表示している。強度は式(28)に従ってバンド幅から変換した値である。

二つのフレームがそれぞれ6つのフォルマントを持つため、格子点は(1, 1)から(6, 6)まで36個になるが、さらにもう一つ(7, 7)という点を設けてある。点(1, 1)から点(7, 7)に向かって格子点をたどりながら進むものとする。例えば図17に示すように(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (7, 7)と進むパスが考えられる。この時、(1, 1)は Frame A の F1 と Frame B の F1 に対応する。(2, 2)以降も同様である。したがって、このようなパスをたどった場合、F1 から F6 の 6 つのフォルマントは全て同じ番号のもの同士が接続されたことになる。しかし、例えば図18に示すように(1, 1), (2, 3), (3, 4), (5, 5), (6, 6), (7, 7)というパスを通ることも考えられる。この場合は Frame A の F2 と Frame B の F3 とが接続され、Frame A の F3 と Frame B の F4 とが接続されることを意味する。Frame A の F4 および Frame B の F2 は接続される相手が存在しない。Frame A の F4 は消えていくフォルマント、Frame B の F2 は出現しつつあるフォルマントとみなされる。

このようにパスの選ばれ方によってフォルマントの接続が決定する。パスの選び方はフォルマント周波数とバンド幅の値の距離によるコスト、および格子点から格子点への移動によるコストを小さくする方法を用いる。

まず、図19に示すように移動に制約を設ける。すなわち、点(i, j)に移動することができるのは(i-1, j-1), (i-2, j-1), (i-1, j-2), (i-2, j-2)の 4 点のみとする。(i-1, j-1)からの移動を A、(i-2, j-1)からの移動を B、(i-1, j-2)からの移動を C、(i-2, j-2)からの移動を D と呼ぶことにする。この制約により、全ての格子点の内、(1, 1)から移動を開始して(7, 7)にたどり着けるために通り得る格子点は明らかに図20に示すものに限られる。

以下、パス探索の手順を図21を参照しながら説明する。

＜ステップ S1＞

まず、Frame A、Frame B のフォルマント個数をそれぞれNA、NBと置く。サイズ NA×NB の配列 C と、サイズ(NA+1)×(NB+1)の配列 ni および nj を用意し、それらの要素を全て 0 で初期化する。C の要素 C(i, j)は、点(i, j)における累積コストを記憶するために用いられる。また、ni の要素 ni(i, j)および nj の要素 nj(i, j)は点(i, j)に最小累積コストで移動してきたパスすなわち点(i, j)への最適パスを記憶するために用いられる。すなわち、点(i, j)への最適パスにおける点(i, j)の直前の点を点(m, n)とした時、ni(i, j)=m、nj(i, j)=n である。

＜ステップ S2＞

全ての可能な点(図20参照)について累積コストと最適パスを計算する。

カウンタ i および j をともに1で初期化する。i、j はそれぞれ Frame A 軸, Frame B 軸のインデックスとして用いられる。

＜ステップ S3＞

点(i, j)に移動可能な4つの点(m, n)についてコスト計算を行う(図19参照)。

カウンタmおよびnを用意し、m = i – 2, n = j – 2と初期化する。また、最小累積コストを求めるためのCminを用意し、可能な限り大きな値を代入しておく。

<ステップS4>

点(m, n)が図20で示す通り得る点の集合に含まれていなければステップS8へ、含まれていればステップS5へすすむ。

<ステップS5>

累積コストを一時的に記憶するCtempを用意し、点(m, n)から点(i, j)へのパスコストと点(m, n)の累積コストの和を記憶する。

<ステップS6>

CtempがCminより小さければ(Yes)ステップS7へ、小さくなければ(No)ステップS8へすすむ。

<ステップS7>

CminにCtempを代入し、ni(i, j)にmを、nj(i, j)にnを格納する。ni(i, j)は点(i, j)に最小累積コストで移動した点のFrame A軸座標、nj(i, j)はFrame B軸座標を記憶する。

<ステップS8>

n = j – 1であれば(Yes)ステップS10へ、n = j – 1でなければ(No)ステップS9へすすむ。

<ステップS9>

nを一つ増加させてステップS4へもどる。

<ステップS10>

m = i – 1であれば(Yes)ステップS12へ、m = i – 1でなければ(No)ステップS11へすすむ。

<ステップS11>

nをj – 2に戻し、mを一つ増加させてステップS4へもどる。

<ステップS12>

iがNA + 1に達していれば(Yes)終了し、達していなければ(No)ステップS13へすすむ。

<ステップS13>

C(i, j)に累積コストを記憶する。すなわち、点(i, j)におけるフォルマント距離(式(26)で計算した値)とCminの和を記憶する。ただし、点(1, 1)はパスの始点であるため、パスコストは存在せず、フォルマント距離のみを記憶する。

<ステップS14>

jがNBに達していれば(Yes)ステップS16へ、達していなければ(No)ステップS15へすすむ。

<ステップS15>

jを一つ増加させてステップS3へもどる。

<ステップS16>

iがNAに達していれば(Yes)ステップS18へ、達していなければ(No)ステップS17へすすむ。

<ステップS17>

jを1に戻し、iを一つ増加させてステップS3へもどる。

<ステップS18>

最後に終点(NA + 1, NB + 1)に最小累積コストで移動する点を求める。

i = NA + 1, j = NB + 1としてステップ S3 へもどる。

パスコストの計算は次のように行う。許されるパスは図19に示す A, B, C, D の 4 つである。Frame A の第 i フォルマントを FA(i)、Frame B の第 j フォルマントを FB(j)で表すと、パス A の場合、FA(i – 1)と FA(i)はそれぞれ FB(j – 1)と FB(j)は互いに接続され、接続されないフォルマントは存在しない。このためパスコスト(言い換えれば disconnection cost)は0 となる。パス B の場合、FA(i – 1)は接続される相手が存在しない。このような場合、パスコストは式(27)に FA(i – 1)の強度を代入することにより計算される。パス C の場合は逆に FB(j – 1)は接続される相手が存在しない。したがって、パスコストは式(27)に FB(j – 1)の強度を代入することにより計算される。パス D の場合は、FA(i – 1)と FB(j – 1)の二つについて接続相手が存在しない。そこで、パスコストは式(27)に FA(i – 1)の強度を代入したものと、FB(j – 1)の強度を代入したものの和となる。

このような計算により、実際のコストがどのようになるかを示す。

図22は、点(i, j)とそれに対して移動可能な 4 つの点(i – 1, j – 1)、(i – 2, j – 1), (i – 1, j – 2), (i – 2, j – 2)を示している。矢印は 4 つの点から(i, j)に対する移動を表しており、矢印の先端には図19で定義したパスの名前 A, B, C, D が表されている。また、4 つの点を表す丸の中には各点における累積コストが記入されている。

パスを表す矢印の中ほどに表した四角で囲んだ数字はパスコストである。例えば、パス B のパスコストはこの移動により接続相手がなくなった Frame A の F3 の強度を用いて式(27)により計算され、11となる。

4 つのパスを通って点(i, j)に達する時の累積コスト(ステップ S5 で計算される Ctemp)は各パスの矢印の終端付近に記入されている。すなわち、移動元の点における累積コストに移動によるパスコストを足した値である。

その結果、パス A は 4035、パス B は 483、パス C は 5351、パス D は 1179 の累積コストを与え、最も小さい累積コストとなるパス B が選択される(ステップ S7)。図23にパス B が選択された様子を示す。パス B が選択されたことにより、ni(i, j)にはパス B の始点の i 軸座標値、nj(i, j)には j 軸座標値が記憶される。また、点(i, j)には、パス B による累積コストと点(i, j)におけるフォルマント距離を式(26)により計算した値 182 を加えた累積コスト 665 が記入された(ステップ S13)。

このようにしてコストを計算しながら最適パスを逐次求め、点(1, 1)から点(NA + 1, NB + 1)まで繰り返す。この後、ni と ni を終点から逆にたどっていけば点(1, 1)から点(NA + 1, NB + 1　　)への最適パスを求めることができる。図24に求められた最適パスを示す。また、その結果、図15に示したフォルマントが接続された様子は図25に示すようになる。Frame A の F1 と Frame B の F1のように互いに接続されるものはフォルマントフィルタを時間的に滑らかに変化させる。また、Frame A の F2 は接続される相手が存在しないため、フォルマントフィルタの中心周波数は変化させずに、強度を徐々に 0 に変化させることで滑らかに消滅させる。逆に、Frame B の F2 は強度を 0 から徐々に大きくしていくことで滑らかに出現させる。

強度を滑らかに変化させるには Fi を一定の速度で変化させる。式(28)を Fb について解くことにより

$$F_b(n)=\begin{cases} -\frac{F_s}{\pi}\log\left(\frac{10^{F_i(n)/20}-1}{10^{F_i(n)/20}+1}\right), & if\ formant \\ -\frac{F_s}{\pi}\log\left(\frac{1-10^{F_i(n)/20}}{1+10^{F_i(n)/20}}\right), & if\ anti-formant \end{cases}$$

が得られる。この式を用いてFiをFbに変換してフィルタ係数を計算すれば良い。

以上のように第5の実施形態による音声合成システムでは、DPマッチングを用いてフォルマントの最適接続を行うため、消えていくフォルマント、現れてくるフォルマントを適切に表現することができる。

（第6の実施形態）

第5の実施形態で説明したようにフォルマントを消滅させたり出現させたりすることにより、フレーム毎にフォルマントフィルタの再割り当てが必要になる。図26は、図25に示したFrame AおよびFrame Bの周辺を表している。また、簡単のために F1～F3 近辺のみを表示している。この図に表示された4つの連続するフレームはFrame AとFrame Bが図25と同じものである。そして、それらをはさむ二つのフレームがFrame AA、Frame BBとして表示されている。Frame AとFrame Bの間では第5の実施形態で述べた方法によりF2とF3が接続されない。図26ではこのことを×印で表現している。接続されないフォルマントは、周波数が同じで強度が非常に弱いフォルマントに向かって消滅するか、逆に出現すると解釈する。

このことを実現するために、図27に示すように、接続相手がないフォルマントに対してはバンド幅が無限大（すなわち強度が 0）のフォルマントを接続相手とする。図27において黒い丸で表したものがそうである。このようにすることで、Frame AとFrame Bの間はフォルマント周波数とバンド幅を補間しながらフィルタを滑らかに変化させ、所望のスペクトルを実現できる。

しかし、Frame AAとFrame Aの間はフォルマントの個数が異なるため、単純な補間では実現できない。Frame AAとFrame BBは図28（a）に示すように3つのフィルタの縦続接続で実現可能である。フォルマントフィルタをこの図では左から順に FF1、FF2 のように表している。しかし、Frame A と Frame B は 5 つのフィルタを縦続接続しなくてはならない。仮に F1 同士が接続されなかったケースを考えると最大で6つのフィルタを縦続接続することになる。図28（b）は6つのフィルタを縦続接続した状態である。

ここでは簡単のためにフォルマントフィルタとして 2 次の単極フィルタ(mono-pole filter)としている。図28上部にはそのフィルタの内部を拡大表示している。D1 および D2 は遅延素子で、1 ステップの値を記憶する。伝達関数は以下のようになる。

$$h(z)=\frac{a}{1+bz^{-1}+cz^{-2}}$$

Frame AA の F1 はそのまま Frame A の F1 になるが、Frame AA の F2 は Frame A の F3 に接続されるため、フィルタの割り当ても考慮が必要である。そこで、常にフィルタを6つ縦続接続しておくこととし、Frame AA から Frame A にかけては以下のような処理を行う。

(1) Frame AA ではフィルタが 3 つしか要らないので、FF4～FF6 は D1 および D2 を 0 にクリアし、a=1, b=0, c=0 とする。こうすれば、フィルタがバイパスされたのと等価な状態にできる。FF1、FF2、および FF3 は a, b, c の値をそれぞれ F1、F2、F3 の周波数とバンド幅から計算する。

(2) Frame AA と Frame A の間は、接続されたフォルマントの軌跡に従って周波数とバンド幅を逐次計算し、滑らかにフィルタ特性を変化させる。

(3) Frame A に達した時点でフォルマントフィルタの割り当て変更を行う。それまでの FF1 は引き続き Frame A における F1 を受け持つ。一方、FF2 は Frame A の F2 を受け持つ。しかし、Frame AA の F2 は Frame A の時点では F3 になる。Frame A の段階で FF2 を F2 に割り当てるとフィルタ係数が急峻に変化するため、クリックノイズが発生する。そこで、それまでの FF2 の係数 a, b, c および内部状態である D1, D2 の値を FF3 にコピーし、FF2 は新たに出現した F2 に割り当てを行う。

このような動作を図29を参照してより具体的に説明する。

図29は、Frame AA, Frame A, Frame B, Frame BB のフォルマントフィルタの構成の変化を表している。各フォルマントフィルタの項目には数字が 3 つずつ表示されている。これら 3 つの数字はそれぞれフォルマント周波数、バンド幅、およびこのフォルマントフィルタに接続された直前フレームのフォルマントフィルタの番号(接続番号)を表している。

例えば、Frame A の FF1 の接続番号には 1 が入っている。これは、Frame AA の FF1 がそのまま Frame A の FF1 に接続されたことを意味する。ところが、Frame A の FF3 には接続番号として 3 ではなく 2 が入っている。これは Frame AA の FF2 が Frame A の FF3 に接続されたことを意味している。また、Frame A の FF2 の接続番号には 0 が入っているが、これは Frame AA から接続されるフィルタがなく、Frame A で新規に出現したフォルマントであることを表している。また、Frame BB には接続番号に 3 が入ったものが存在しない。このことは Frame B の F3 はその時点で接続相手が存在せず、消滅したことを意味する。3 つの数字が全て 0 のものは、フィルタとしての機能が不要で、バイパスされる、すなわち係数の値が a=1, b=0, c=0 であることを意味する。

さて、Frame AA から Frame A に状態が変化する時、フィルタの再割り当てを図30に示す手順で行う。

FF6 から順に FF1 に向かって繰り返し(ステップ S31～ステップ S39)

if 接続番号が 0(ステップ S32)

D1、D2 をクリア(ステップ S33)

else

接続番号を N とすると N 番目のフォルマントフィルタ FFN から D1、D2 をコピー(ステップ S34)

endif

フォルマント周波数とバンド幅から a, b, c を計算してセットする(ステップ S36)。ただし、フォルマント周波数とバンド幅がともに 0 の場合は a=1, b=0, c=0 とする(ステップ S37)。

繰り返し終了

以上のように第6の実施形態による音声合成システムでは、DP マッチングによるフォルマント最適接続結果に従ってフィルタの縦続接続の構成を変更する仕組みを持つため、DP マッチングにより最適に接続されたフォルマントに従ったスペクトルを滑らかに再現することが可能となり、クリックノイズや波形上の不連続点の発生を防ぎ、滑らかな音声を合成することができる。

JP05-297900

クレーム

1．音声合成システムは、

音声生成モデルに基づき推定されたフォルマントパラメータ(フォルマント周波数とフォルマントバンド幅とを含む)の時系列データを利用して音声を合成するものであり、

隣り合ったフレーム間でのフォルマントパラメータの対応関係を動的計画法(dynamic programming)を用いて決定する。

2．クレーム1に記載の音声合成システムは、

フォルマントパラメータの対応関係の決定において、αとβを所定の重み係数、$F_f(n)$を第 n フレームのフォルマント周波数、$F_i(n)$を第 n フレームのフォルマント強度、ε を所定の値とするとき、

$$d_c(F(n),F(n+1)) = \alpha\left|F_f(n) - F_f(n+1)\right| + \beta\left|F_i(n) - F_i(n+1)\right|$$

$$d_d(F(k)) = \alpha\left|F_f(k) - F_f(k)\right| + \beta\left|F_i(k) - \varepsilon\right|$$

$$= \beta\left|F_i(k) - \varepsilon\right|$$

によって接続コスト$d_c(F(n), F(n+1))$および非接続コスト$d_d(F(k))$を求め、動的計画法における格子点移動のコストに用いる。

3．クレーム2に記載の音声合成システムは、

互いに接続されないフォルマントが含まれる隣り合ったフレームにおいて、

接続される相手が存在しないフォルマントと同じ周波数で強度が 0 のフォルマントをもう一方のフレームに配置し、

二つのフレームの間をフォルマント周波数と強度を滑らかな関数に従って補間して接続する。

4．クレーム2に記載の音声合成システムは、

$F_b(n)$を第 n フレームのフォルマントバンド幅、F_sをサンプリング周波数とするとき、

$$F_i(n) = \begin{cases} 20\log_{10}\left(\dfrac{1+e^{-\pi F_b(n)/F_s}}{1-e^{-\pi F_b(n)/F_s}}\right) & \text{, if formant} \\ 20\log_{10}\left(\dfrac{1-e^{-\pi F_b(n)/F_s}}{1+e^{-\pi F_b(n)/F_s}}\right) & \text{, if anti - formant} \end{cases}$$

によってフォルマント強度 $F_i(n)$を計算する。

5．クレーム3に記載の音声合成システムは、

複数のフォルマントを含む声道伝達関数を複数のフィルタの縦続接続によって実現し、

隣り合うフレームの間で接続が行われないフォルマントがあることによりフィルタの接続を変更する必要がある場合に、

フィルタの係数および内部記憶値を別のフィルタにコピーし、

自身のフィルタの係数および内部記憶値は別のフィルタからコピーするか所定の値に初期化す

る。

6. クレーム4に記載の音声合成システムは、
複数のフォルマントを含む声道伝達関数を複数のフィルタの縦続接続によって実現し、
隣り合うフレームの間で接続が行われないフォルマントがあることによりフィルタの接続を変更する必要がある場合に、
フィルタの係数および内部記憶値を別のフィルタにコピーし、
自身のフィルタの係数および内部記憶値は別のフィルタからコピーするか所定の値に初期化する。

7. 音声分析方法は、RK 音源モデルなどの声帯音源モデルを利用して音声信号波形の音源パラメータと声道パラメータを推定するものであり、
推定された声道伝達関数の逆特性にて構成されたフィルタを用いて推定音源波形を抽出し、
前記推定音源波形の声門閉鎖タイミング(GCI: glottal closure instance)に対応するピーク位置を2次関数などのあてはめによってサンプリング周期よりも高い時間精度で推定し、
前記推定されたピーク位置の近傍のサンプル位置に GCI を同期させて音源モデル波形を生成し、
前記生成された音源モデル波形をオールパスフィルタでサンプル周期よりも高い時間精度で時間的に移動することにより GCI を前記推定されたピーク位置に一致させる。

8. 音声分析方法は、RK 音源モデルまたはその拡張として定義される声帯音源モデルを利用して音声信号波形の音源パラメータと声道パラメータを推定するものであり、
推定された声道伝達関数の逆特性にて構成されたフィルタを用いて推定音源波形を抽出し、
前記推定音源波形の DFT(discrete Fourier transformation)における基本波レベルをH1、第 2 高調波レベルを H2 として、HD=H2-H1 で定義される HD の値から声門開放時間率(open quotient)OQ を推定する。

9. クレーム8に記載の音声分析方法は、
OQ の推定に
$$OQ = 3.65HD - 0.273HD^2 + 0.0224HD^3 + 50.7$$
の関係を用いる。

開示の概要

分析対象音声波形は複数ピッチ周期を含む長さの窓で切り出され、RKモデル音源パラメータ推定が行われる。また複数の音源パルスについてGCI(glottal closure instance)をすべて推定する。これらに基づいてRKモデル音源波形を生成し、音声波形との関係をARXシステム同定によって解析し、声道伝達関数を推定する。この処理を繰り返しながら GCI が一定値に収束したときに同定を完了する。これにより、音声信号の音源パラメータと声道パラメータを精度よく分離し高品質な分析合成システムを実現することができる。

JP2005070430A 20050317